Hindawi
International Transactions on Electrical Energy Systems
Volume 2023, Article ID 1190103, 15 pages
https://doi.org/10.1155/2023/1190103

Research Article

Decentralized Reinforcement Learning Approach for Microgrid Energy Management in Stochastic Environment

Razieh Darshi,[1] **Saeed Shamaghdari**,[1] **Aliakbar Jalali**,[1] **and Hamidreza Arasteh**[2]

[1]*School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran*
[2]*Power Systems Operation and Planning Research Department, Niroo Research Institute, Tehran, Iran*

Correspondence should be addressed to Saeed Shamaghdari; shamaghdari@iust.ac.ir

Received 25 October 2022; Revised 12 January 2023; Accepted 17 January 2023; Published 11 February 2023

Academic Editor: Salvatore Favuzza

Microgrids are considered to be smart power grids that can integrate Distributed Energy Resources (DERs) in the main grid cleanly and reliably. Due to the random and unpredictable nature of Renewable Energy Sources (RESs) and electricity demand, designing a control system for microgrid energy management is a complex task. In addition, the policies of microgrid agents are changing over time to improve their expected profits. Therefore, the problem is stochastic and the policies of the agents are not stationary and deterministic. This paper proposes a fully decentralized multiagent Energy Management System (EMS) for microgrids using the reinforcement learning and stochastic game. The microgrid agents, comprising customers, and DERs are considered as intelligent and autonomous decision makers. The proposed method solves a distributed optimization problem for each self-interested decision maker. Interactions between the decision makers and the environment during the learning phase lead the system to converge to the optimal equilibrium point in which the benefits of all the agents are maximized. Simulation studies using a real dataset demonstrate the effectiveness of the proposed method for the hourly energy management of microgrids.

1. Introduction

DERs are emerging as the recently developed technologies for supplying the growing demand for electricity and thermal energy [1]. DERs have lower environmental impacts than traditional energy sources, such as petroleum, natural gas, and coal. Recently, the utilization of DERs has attracted a great deal of attention due to their environmental, economic, and technical advantages [2]. The secure supply of electricity based on DERs is a reliable, efficient, and environmentally friendly replacement for the conventional centralized energy sources [3]. Distributed energy sources include renewable energies, nonrenewable energies, and batteries. It is complicate to combine the RESs directly into microgrids due to their random and intermittence features affected by the meteorological parameters. Microgrids are the interface between the utility grid and distributed RESs. They comprise DERs, storage systems, and local loads, as well as control systems and different entities (such as microgrid operators). Microgrids are defined as local and small distribution systems, which include load and generation units [4]. They can operate in grid-connected and islanded modes [5]. Microgrids should guarantee various functions for instance supply of electrical and thermal demand, involvement in the energy market, maximization of customers' satisfaction level, and optimization of generators' benefit [6]. Several research studies in the field of microgrids control have been presented in the literature to supply the abovementioned functions of the microgrids. The design of a multiagent EMS for microgrids is a complicate task due to the various range of required functions. In a typical microgrid, there are different types of agents including generators, customers, and energy storage systems. An efficient EMS should be designed in a way that in addition to supplying all the demands and reducing the consumers' expenses, it maximizes the producers' profits. Therefore, it is necessary to design a multiagent system for microgrid energy management to maximize the profit of all the agents. In a microgrid with multi-intelligent agents, the profit of each agent relies on its actions and also the actions of other

agents. The policy of each agent is not deterministic and stationary, since each agent can change its policy over time to maximize the expected profit during the learning process. Therefore, the environment is stochastic from the point of view of each agent.

Many of the existing methods for microgrid energy management utilize a centralized structure. In [7], a dynamic EMS has been developed using the adaptive dynamic programming approach, in which the critical loads were supplied at all times. A home EMS based on Levenberg–Marquardt algorithm has been presented to optimize the customers' performance in presence of a grid-connected photovoltaic system with a battery energy storage system in [8]. An energy management structure based on the Nash Q-learning algorithm has been used in [9] to manage all the DERs. In the Nash Q-learning algorithm [10], each agent observes all the information including its reward, and also the actions and rewards of other agents. In practice, implementing this method is more complicated than central methods. Against the central systems where information is available for one unit, in the Nash Q-learning approach, all the agents are aware of all the information. A deep reinforcement learning method has been presented in [11] to minimize the daily operating cost of a microgrid. Using the deep feedforward neural network, the optimal action-value function is approximated. Various deep reinforcement learning methods were presented in [12] to design microgrid energy management. For the implementation of the EMS using deep reinforcement learning, there is a centralized controller. Therefore, information from all other agents is expected to be available for the central system. So, such a method is faced with high communication complexity and computational cost [13].

In [14], a hierarchical reinforcement learning-based approach has been developed to achieve an optimal policy of the microgrid EMS. Instead of receiving information from neighboring agents, a state variable "Trend" was introduced to solve the dimensionality issue of the multiagent system. The hierarchical reinforcement learning algorithm converges to a locally optimal policy, named recursively optimal policy [15].

Several studies have used the decentralized method for energy scheduling of microgrids. In [16], a fully decentralized method based on a reinforcement learning algorithm and multiplayer games has been established. Each decision maker aims to maximize its own long-term profit. In this method, each player only observes its actions, but the problem is stateless and the agents cannot observe the states of the environment. In [17], the trajectory planning of unmanned aerial vehicles for energy efficiency maximization has been proposed using a decentralized learning-based approach. But in the Q-function, the action of all the other agents should be available. Energy management strategy based on decentralized fuzzy logic controller has been extended for charging of electric vehicles in [18]. In [19], a decentralized EMS for autonomous polygeneration microgrid has been designed; however, each agent communicated with each other to calculate optimal control. Energy management problem has been modeled by cooperative and noncooperative game theory and the agents are allowed to communicate in cooperative game in [20]. Due to the communication of agents with each other, the computational and communication complexity has increased in these methods. Besides, in practice, agents prefer not to share their information to maximize their profit in the competitive electricity market. Some articles use conventional Q-learning of single agent for decentralized multiagent structure. In [21], distributed energy scheduling of a microgrid has been developed based on a multiagent model and conventional Q-learning. Every customer and supplier do not have access to other agents' information and prior information about the environment. A multiagent structure for home energy management is proposed in [22] using single agent Q-learning and neural networks. By the scheduling of the household appliances and electric vehicles, the electricity bill has been decreased, but the profit of the DERs has not been considered. They have used normal single agent Q-learning in a decentralized structure to find the optimal solution. The single agent Q-learning are developed for stationary systems and no guarantee is provided for its convergence to the optimal solution in stochastic systems. For a special case, the nonconvergence of Q-learning in Shapley's game has been shown, see Section 4 in [23]. Finding the optimal policy in multiagent systems is complicated due to the nonstationary environment. Each agent tries to learn its optimal policy; however, the policies of the agents are changing over the time to maximize their expected profits.

To the best of our knowledge, despite many research on the multiagent EMS of microgrids, none of the proposed methods offers a fully decentralized reinforcement algorithm consistent with the stochastic game structure of the microgrid EMS, so that there is a guarantee of convergence to the optimal solution in the stochastic environment. Therefore, in this paper, a model-free Q-learning algorithm is developed to control the multiagent EMS of microgrids in the stochastic environment based on the stochastic game and reinforcement learning. The multiagent EMS is modeled using the Markov game [24]. In this method, the profit of all the DERs, including renewable and nonrenewable units and battery energy storage systems are maximized. Meanwhile, the expenses of all the customers are minimized. Indeed, the benefits of all the agents of the microgrid are optimized, concurrently. Each agent receives only its reward from the environment to learn the consequences of its actions. It is not aware of the actions and even the existence of other agents. The proposed algorithm converges to a suboptimal solution. The main innovations of this paper are summarized as follows:

(1) The problem of microgrid EMS is modeled using the stochastic dynamic games.

(2) A fully decentralized Q-learning algorithm applicable to the stochastic game of EMS is developed.

(3) All the customers and energy generators are considered as intelligent and independent agents. These agents can make decisions to maximize their profits.

(4) Each agent learns its optimal policy with minimal information, while the computational cost and communication complexity are reduced.

The rest of this paper is organized as follows: Section 2 describes the structure of the microgrid. The Microgrid EMS scheme using a model-free reinforcement learning algorithm is developed in Section 3. Finally, the simulation results and concluding remarks are presented in Sections 4 and 5, respectively.

2. Problem Description

Microgrids as small-scale and low-voltage power grids are connected to the main grid through a common connection point called the point of common coupling. In the grid-connected mode, microgrids satisfy supply and demand balance by selling/purchasing the excess/deficit energy to/from the main grid. However, the reduction of the microgrid dependency on the main grid is a main goal in the microgrid energy management problem. Therefore, in addition to increasing the profit of all the agents in the microgrid, the EMS of microgrids should be designed in a way that the dependence of the microgrid on the main grid is reduced [25]. In this paper, the grid-connected mode of the microgrid is considered. Loads in microgrids are divided into two categories, controllable and noncontrollable. Uncontrollable loads such as medical center systems and essential tasks in the industry must be provided at the time of demand. These loads are inflexible to time and cannot be moved over time. However, controllable loads can be removed or transferred to low-load times. Figure 1 shows the structure of a microgrid consisting of solar panels, wind turbine, diesel generator, electric and thermal fuel cell, electric and thermal microturbine, battery, and several local electric and thermal loads. A microgrid operator is considered as a high-level controller in power microgrids.

In order to guarantee the reliability and security of a network, the required power must be supplied by the producers at all times. In the connected mode to the main grid, the constraint of power balance means the equality of the generated power with the consumed loads [26]. Therefore, for electric loads, the power balance constraint is defined as follows:

$$\sum_{i=1}^{n} \text{Load}_i^E = P_w + P_{PV} + P_d + P_b + P_{MT}^E + P_{FC}^E + P_{\text{main}}^E, \quad (1)$$

where Load_i^E is the amount of electric load demand of the i^{th} consumption agent and n is the number of electric consumer agents. P_w, P_{PV}, P_d, P_b,P_{MT}^H, P_{FC}^H, and P_{main}^E are the electrical power output of the wind turbine, photovoltaic solar panels, diesel generator, battery, microturbine, fuel cell, and main grid. The power balance condition for thermal loads is also defined as follows:

$$\sum_{i=1}^{m} \text{Load}_i^H = P_{MT}^H + P_{FC}^H + P_{\text{main}}^H, \quad (2)$$

where Load_i^H is the heat load demand of the i^{th} consumer agent and m is the number of thermal consumer agents. P_{MT}^H, P_{FC}^H, and P_{main}^H are the generated thermal power of microturbine, fuel cell, and main gird, respectively.

The capacity constraints represent the operating range of distributed generators and have the following range:

$$P_i^{\min} < P_i(t) < P_i^{\max}, \quad (3)$$

where the output power of distributed generator i in the time interval t is determined by $P_i(t)$. $P_i^{\min}$ and $P_i^{\max}$ are the minimum and maximum output power of the generator i, respectively.

SOC indicates the state of charge of the battery. The following technical constraint is applied to prevent excessive charging and discharging of the battery energy storage system:

$$\text{SOC}_{\min} \geq \text{SOC}(t) \geq \text{SOC}_{\max}, \quad (4)$$

where $\text{SOC}_{\min}$ and $\text{SOC}_{\max}$ are the minimum and maximum charge levels of the battery. In this research, SOC is limited to the range of [0.2, 0.8] to avoid the damage of the battery.

3. Microgrid EMS Based on Decentralized Reinforcement Learning

Reinforcement learning is an action-based learning. In this method, an agent tries to improve its actions and control policies by influencing and receiving better feedback from the environment. Indeed, the agent tries to correct its actions by receiving rewards and not punishing in interaction with the environment. Correct control decisions should be retained in the system memory by the reinforcement signal so that they are more likely to be used next time [27].

Reinforcement learning uses the formal structure of Markov decision processes and describes the relationship between a learning agent and the environment using states, actions, and rewards. In each time interval t, the reinforcement learning agent can observe the states of the environment, S_t, and perform actions, A_t, based on the observed states. In a later period, as a result of its action, the agent receives a numerical reward, R_{t+1}, and goes to a new state, S_{t+1}. Therefore, by using action and reaction with the environment, an agent learns to choose actions that maximize its reward. A reward is a number calculated using the reward function and is defined according to the purpose of the reinforcement learning problem. The goal of the intelligent agent is to maximize all the rewards received in a long time [28].

The action-value function, $Q_\pi(s, a)$, is the expected value of the sum of weighted rewards (with discount factor) in state s, performed action a, and under policy π. It is expressed as follows:

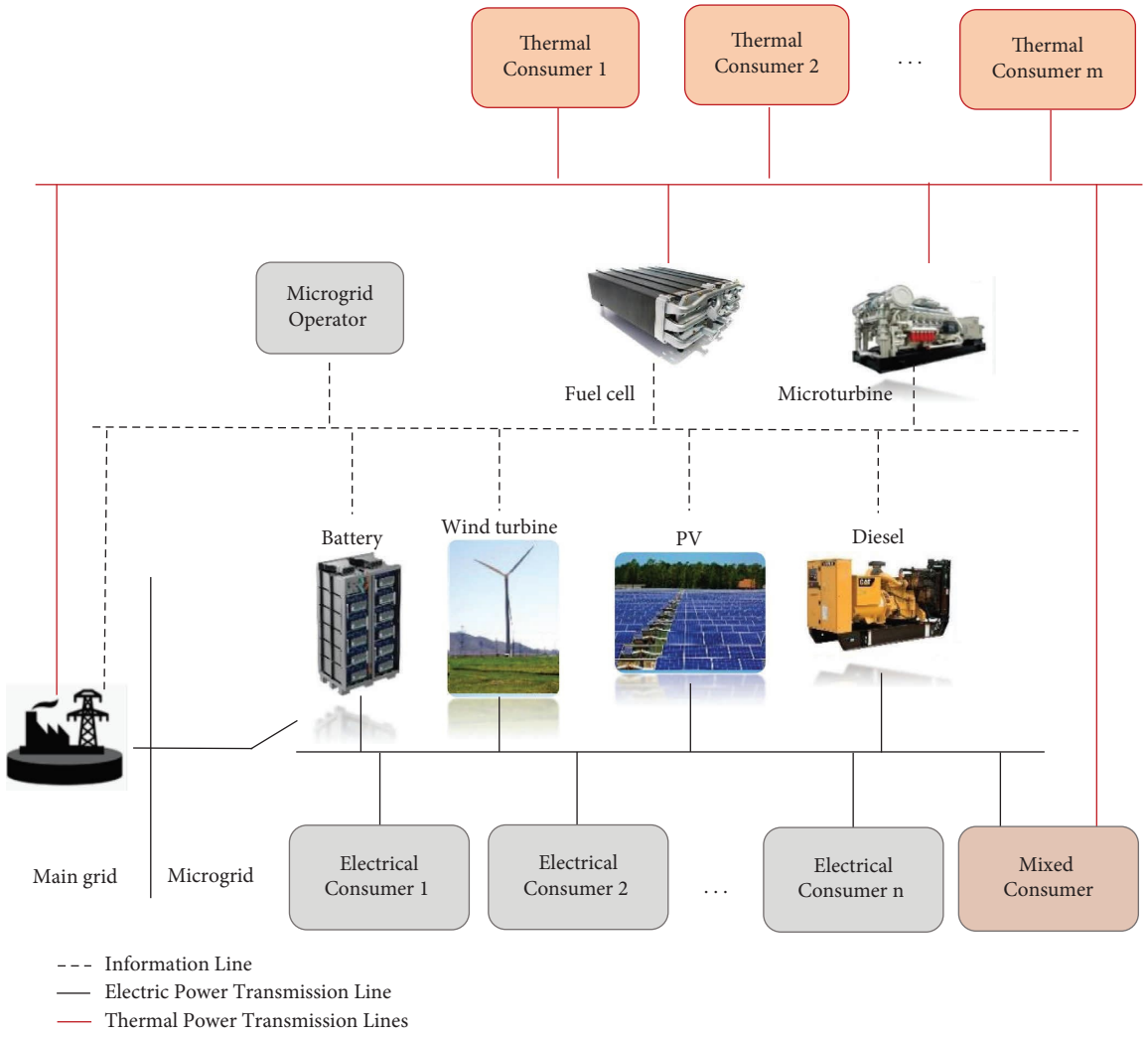

Figure 1: Microgrid structure.

$$Q_\pi(s,a) \doteq E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]. \quad (5)$$

The parameter γ is called the discount rate and has a value between zero and one. This parameter represents the current value of future rewards. When it approaches one, the agent pays great attention to future profits. R_t is the reward function at time t. Even if an accurate and complete model of environment dynamics is available, calculating the optimal policy by solving Bellman's optimality equation is not easily possible. In order to calculate the optimal policy, the Q-learning update law can be used for a single agent system in a stationary environment [29]. There is no guarantee for the standard Q-learning to converge in the stochastic game due to the presence of multiple active learner agents. The environment for all agents becomes a nonstationary as a consequence of the existence of these learning agents. For this reason, in the following, a decentralized Q-learning algorithm based on a stochastic game for the multiagent microgrid EMS is presented, which is in accordance with the stochastic structure of the environment based on [30].

Figure 2 depicts the structure of the decentralized control system. A controller (agent) is interested in maximizing its own long-term reward. Each agent only has access to the state of the environment and its reward. It is not even aware of the existence of the other agents and their information. The state of the environment at time slot t is S_t, the reward and the controller signal (actions) of agent i at time t are A_t^i and R_t^i. W_t is the random disturbances at time slot t.

The element of a finite discounted stochastic game is as follows:

(i) S: the finite set of states

(ii) N: the finite number of the agents

(iii) A^i: the finite set of actions (control decisions) for agent i

(iv) $R^i(s, a^1, \ldots, a^N)$: reward function for calculating the reward of agent i, for all $s \in S$ and, $a^1 \in \mathrm{A}^{\mathrm{i}}$

(v) γ^i: the discount factor of agent i

(vi) s_0: a random initial state belongs to S

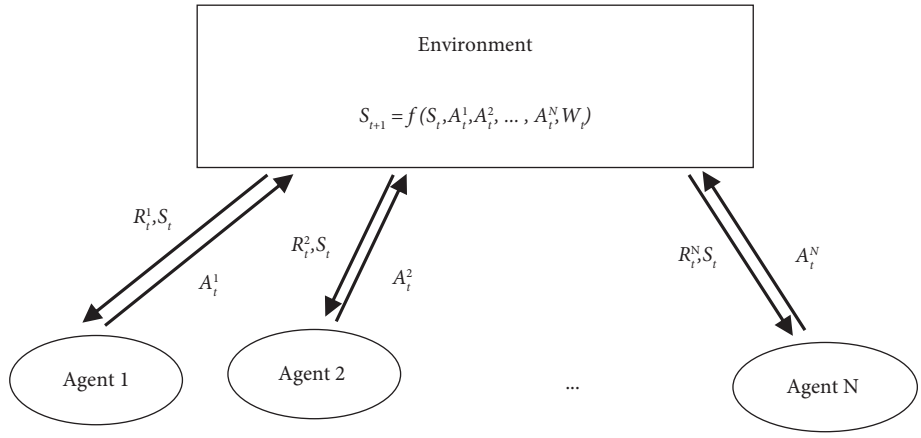

Figure 2: The structure of decentralized control system.

(vii) $(p(s'|s, a^1, \ldots, a^N) \doteq \Pr[S_t = s' S_{t-1} = s, A_{t-1}^i = a^i])$: the conditional probability distribution function, where p: $S \times S \times A \longrightarrow [0, 1]$ and $\forall s', s \in S$ and $\forall a^i \in A^i$

The above stochastic game is a discrete-time Markov process starting with initial state s_0. At each time slot, the agent i choose possibly randomly actions $a^i \in A^i$. The rule of selecting an appropriate action based on the agent's observation history at the time t is called policy. Although the environment is stochastic, the focus of this paper is on stationary policies and agents choose their actions solely based on the state S_t. By introducing a new algorithm for microgrid energy management, the structure of the stochastic game is transformed into a stationary environment at each time slot. It is proved that the problem converges to an equilibrium solution.

In the multiagent environment, the agent's reward relies not only on its actions and the state information but also on the action of other agents. The dynamic of self-interested agents generally is modeled by the framework of stochastic games, which generalized Markov decision problems [31]. Joint actions include actions of all agents taking at each time slot. The state transition of the system is complex due to the joint actions. At the beginning of each time slot, the environment moves into a new stochastic state, which is affected by the previous state and actions of all agents. Therefore, the Markov property is satisfied by the state-action transition. In proposed method, all agents have constant policies at some special phases named exploration periods. The k^{th} exploration phase is implemented during times $t = t_k, \ldots, t_{k+1} - 1$, where $t_{k+1} = t_k + T_k$, as demonstrated in Figure 3.

The length of the exploration period is denoted by $T_k \varepsilon [1, \infty)$. All agents have constant policies during each exploration phase. The critical point is to generate at each exploration period a stationary environment. The agents explore and learn their optimal policies and Q-function at each period corresponding to their consistent policies. Each agent has two Q-functions. During the exploration phase, each agent selects actions based on its Q_0-function but updates its Q_1-function based on its observations. Therefore, the policies of agents do not change during the exploration phase. The Q-learning update rule of agent i in each iteration is defined as follows:

$$Q_1^{L+1,i}\left(s_t, a_t^i\right) = \left(1 - \alpha_t^i\right) Q_1^{L,i}\left(s_t, a_t^i\right) + \alpha_t^i \left[r_{t+1}^i + \gamma^i \max_a Q_1^{L,i}\left(s_{t+1}, a\right)\right], \tag{6}$$

where $\gamma^i, \alpha_t^i \in [0, 1]$ are the discount rate and learning rate of agent i. It is transferred from state s_t to the next state s_{t+1} by doing a_t^i and receives the reward r_{t+1}^i. The Q-learning algorithm is a model-independent reinforcement learning method. At the end of each exploration phase, Q_0 is updated with Q_1 for each agent with the probability $1 - \mu$. The proposed method is summarized in Algorithm 1. If the agents are updated according to Algorithm 1, in each learning phase, the environment becomes a stationary environment. Finally, if the value of T_k is large enough and the convergence conditions of normal Q-learning in [29] are satisfied, the decentralized stochastic game will converge to an optimal or suboptimal solution [30].

The goal of the EMS for a microgrid is to maximize the profit of all agents over a long time. For this reason, the total profit of the i^{th} generator for a long period is defined as follows:

$$\max F_i = \sum_{t=1}^{\infty} \gamma^t * \left[Pr_i(t) \times P_i^{\text{mic}}(t) + S_p(t) \times P_i^{\text{main}}(t) - C_i^{op}\left(P_i^{\text{mic}}(t) + P_i^{\text{main}}(t)\right)\right], \tag{7}$$

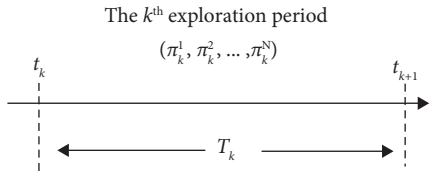

FIGURE 3: The k^{th} exploration period.

where t is the time period. The parameter γ is the discount rate. $P_i^{\text{mic}}(t)$ and $P_i^{\text{main}}(t)$ are the power sold by the i^{th} generator to the microgrid and the main grid at the time t, respectively. $Pr_i(t)$ is the offered price by the i^{th} generator for selling energy to the microgrid. The state of the environment, $S_p(t)$, is the energy purchase price by the main grid from the microgrid. $C_i^{op}(\cdot)$ is the operating cost function of the i^{th} generator. The cost function is calculated practically [32].

The objective function of battery energy storage systems is given as follows:

$$\max F_b = \sum_{t=1}^{\infty} \gamma^t * \left[\left(Pr_b(t) \times P_b^{\text{mic}}(t) + S_p(t) \times P_b^{\text{main}}(t)\right) - \left(Pr_m(t) \times P_b^{\text{input}}(t)\right)\right]. \tag{8}$$

The first term is the profit from selling energy. The second term is the cost of buying energy. In any time, the battery can be a buyer or seller of energy. $Pr_b(t)$ is the offered price for selling energy by the battery. $P_b^{\text{mic}}(t)$ and $P_b^{\text{main}}(t)$ are the power sold by the battery to the microgrid and the main grid in the time interval t, respectively. P_b^{input} and Pr_m are the amount of purchasing power by the battery and the electricity market price. The policy of energy storage systems should be such that energy is purchased at low price times and returned to the grid at the consumption peak. The goals of consumer agents are to minimize costs, which are calculated as follows:

$$\min F_c = \sum_{t=1}^{\infty} \gamma^t \left[\text{Pr}_m(t) \times \left(L_i^{NC}(t) + \beta(t) \times L_i^C(t)\right) + \mu \times (1 - \beta(t)) \times L_i^C(t)\right], \tag{9}$$

where $L_i^{NC}(t)$ and $L_i^C(t)$ are uncontrollable and controllable loads in the time interval t, respectively. $\beta(t)$ is the percentage of controllable load curtailment. μ is the coefficient of consumer dissatisfaction for load curtailment and depends on the type of consumer and their enthusiasm for managing and optimizing their consumption and expenses.

4. Numerical Study

For all the agents, except the battery, the states include (t, S_s, S_p). t is the time slot, S_s is the price of selling energy to the main grid and S_p is the price of buying energy from the main grid. In addition to the above states, the battery agent has an additional state including the battery's SOC, which changes between 0 and 100 percent. The set of actions for microturbine and fuel cell includes the amount of produced electrical power, the amount of produced thermal power, and the bid price for selling energy to the microgrid. The diesel generator also decides on the amount of electrical output power and the bid price. The action set of RESs (including wind turbine and solar panels) only includes the bid price for selling energy. If the bid price of RESs is equal to the price of nonrenewable sources, the priority is with the renewable energy sources. Indeed, microgrid buys the power produced by the wind turbine and solar panel and then uses energy from other sources if needed. If the produced power is more than the requirement of the microgrid; DERs can sell excess energy directly to the main grid. Since the price of selling energy to the main grid is much lower than the price of purchasing energy from microgrid, all generator agents should be trained in a way that they sell their produced power inside the microgrid by offering a reasonable price in the competitive electricity market. Therefore, the microgrids can supply the required power from domestic producers instead of buying from the main grid. As a result, the profit of domestic producers is increased and the dependence of the microgrid on the main grid is also reduced. The action set of the battery includes the state of charge or discharge, the amount of power exchanged and the bid price. In charging mode, the battery power is negative and in discharging mode, it is positive. The energy consumption is a random variable and relies on parameters such as weather condition and consumption time [33]. In order to create dataset for learning the system, the amount of demand is modeled with exponential distribution function. Demand can be divided into uncontrollable and controllable loads. There is no control over the first category and they must be supplied at the time of demand. However, the percentage of curtailment of controllable loads is a decision variable.

Since the goal of the reinforcement learning problem is to maximize the objective functions, the immediate reward is defined in a way that maximizes functions (7)–(9). The reward of DERs is the amount of the net benefit from selling

//initialization
(1) Initialize the learning parameters γ^i and α^i,
(2) Set K_1, T_k, $\square$, μ,
(3) Initializes $Q_0^i, Q_1^i = 0$ for all states and actions.
//learning
(4) **For** $k = 1$: K_1
//exploration period
(5) **For** $n = 1$: T_k
(6) **For** $t = 1$: 24
(7) Each agent senses the states of the environment
(8) The demand is predicted by the exponential random distribution
(9) The outputs of wind and PV are determined.
(10) Each agent takes random actions with the probability 1-$\square$ and selects the best action with the probability $\square$ based on Q_0^i
(11) MO clears the market
(12) Each agent observes its immediate reward
(13) The $Q_1^{L,i}$-function for each agent is updated according to equation (6)
(14) **End**
(15) **End**
//end exploration period
(16) Q_0^i is updated with $Q_1^{T_k,i}$ for each agent with the probability $1-\mu$
(17) **End**
//end learning

Algorithm 1: The proposed energy management system for smart microgrids.

energy. The reward of consumers is the negative value of the electricity bill and their dissatisfaction level.

The EMS for a microgrid has been simulated using actual data of renewable energy output power and data from the IREMA website. The output powers of wind turbine and solar panels have been collected hourly.

The proposed microgrid consists of thermal and electrical energy sources, a battery energy storage system, and electrical and thermal loads (see Figure 1). The specifications of distributed resources are according to Table 1. Due to the environmental pollution of nonrenewable energy resources, the capacity of the microturbine and diesel generator are considered lower than the capacity of renewable resources to limit the use of nonrenewable resources in power grids. Attribute to the high cost of purchasing battery and short battery life, the capacity of the battery is also restricted. Solar and wind energy sources are significantly available in the country; hence, the capacity of wind and solar generators has been designed more than other generators. Four electrical consumer agents, three thermal consumer agents, and one electric and thermal consumer (mixed customer) have been considered in the microgrid with the capacity of 8, 4, and 8 kW, respectively. Due to consumption management, the total capacity of the generators is considered less than the total power of consumers. Consumers can manage up to 70% of their consumption. The remaining is considered as a noncontrollable load, which must be supplied at the time of demand. One day is divided into 24 one-hour periods. In each period, the exchange rate from the main grid is in the range of 150–1200 Rials/kWh. According to Iran's electricity market on the IREMA site, the bid price by DERs has been set between 200–1300 Rials/kWh. The practical data were collected in the summer season. To verify the performance of the proposed method in the summer season, the algorithm has been simulated for 80 days in each scenario. If the simulation is run for more days, the performance of the algorithm will be the same as before. The presented method has been evaluated under two scenarios: without learning and all agents learning. Each scenario has been simulated for 80 days. The total duration of the simulation is 160 days. In the first 80 days, there is no learning and all requested loads are satisfied, and distributed energy sources randomly choose an action. In the second 80 days, all agents are trained and have the ability to make intelligent decisions.

Table 1: The capacity of distributed energy resources.

DER	Wind	PV	BESS	MT	FC	Diesel
P_{rated} (kW)	10	10	5	6	6	5

In the learning period, T_k is equal to 120 days and exploration phase is iterated 1500 times. The evaluation period for each scenario is simulated ten times. The average evaluation results of the proposed EMS are shown in Figures 4–12. The average amount of profit and power for the two scenarios is shown in Table 2. In Table 2, the cost is for one consumer agent and power includes the total requested load in the microgrid. As shown in Figures 4 and 5, although the average output of wind turbine and solar panels in the second scenario (second 80 days) has not changed, their profit has increased significantly, since the resources can make smarter decisions.

In Figures 6–8, the daily average profit and output power of diesel generator, fuel cell, and microturbine are shown. According to the training of the generator agents in the second scenario, the profit of the diesel generator, fuel cell,

Table 2: The average results of EMS based on decentralized reinforcement learning over 800 days.

Scenario	Daily generated/consumed energy (kWh)		Daily profit/expense (rial)	
	I	II	I	II
Wind	125.2	124.2	40561	51681
PV	70.2	70.1	31226	41279
Diesel	63	117.4	14670	28274
Fuel cell	75.2	141.7	17153	41539
Microturbine	75.3	142.1	24401	52317
Battery	10.7	1.2	−2312	91
Electrical load	618.9	216.6	81184	28049
Thermal load	257.1	99	60282	22849
Main grid	475.7	−211	436720	−46247

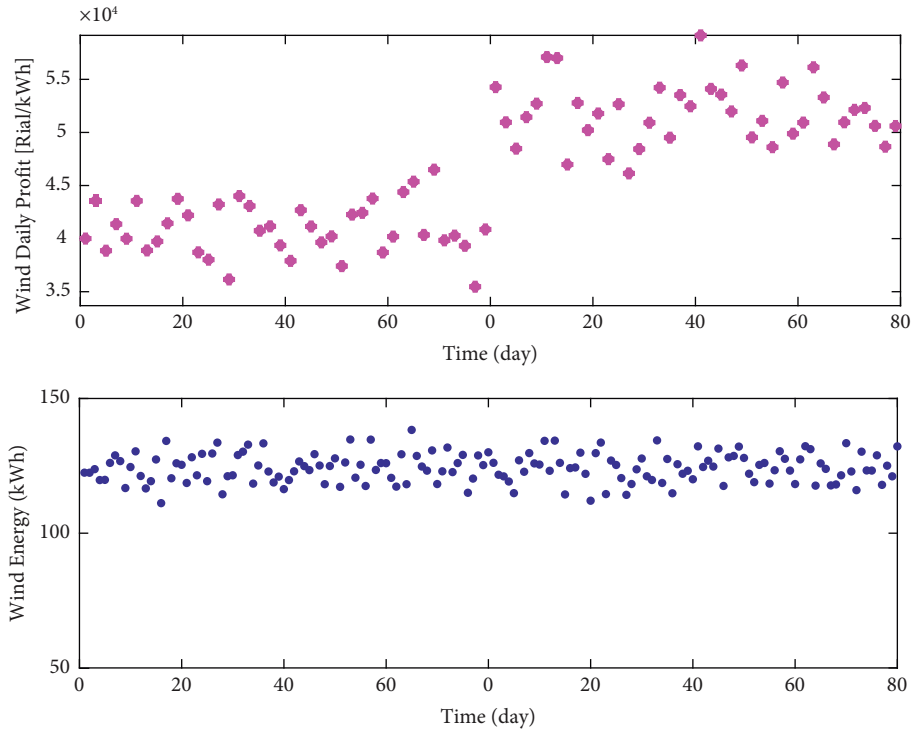

Figure 4: Average daily profit and output power of wind turbine.

and microturbine agents has increased. The ratios of profit to production in diesel generator in the first and second scenario are 232.9 and 240.8, respectively. Therefore, although production has increased in the second scenario, the ratio of profit to production (according to Table 2) has also increased for diesel generators. The diesel generator can intelligently transfer its production to the hours when the demand and energy cost are high. Indeed, this agent has learned to sell more energy inside the microgrid and its profit has increased by offering a reasonable price to sell energy. The ratios of profit to production in the first and second scenarios are 228.1 and 293.1 for the fuel cell and 324.1 and 368.2 for the microturbine, respectively. Therefore, just like the diesel generator, these agents can make more optimal decisions by exploring and exploiting the environment during training.

Figure 9 shows the simulation results for the battery. In the second scenario, where the battery is trained, its profit is positive. At other times, it is negative. A negative profit means that the battery had bought energy at a high cost and sold it when the price of electricity was low.

Figures 10 and 11 show the results of electrical and thermal consumers, respectively. To compare fairly the scenarios, the ratios of cost to consumption in the first and second scenarios have been compared. For electrical consumers, the ratios are 131.2 and 129.5 in the first and second scenarios, respectively. The reduction of these ratios illustrates that the consumer agents have been able to manage and transfer their consumption to low price time. The agent reduces its consumption when the price of electricity is high and increases it when the price is low. The dissatisfaction coefficient (μ) is 10. By adjusting the μ parameter, the agents trade off between cost and comfort. According to Table 2, it can be seen that the above result is also true for the thermal consumer (see Figure 11).

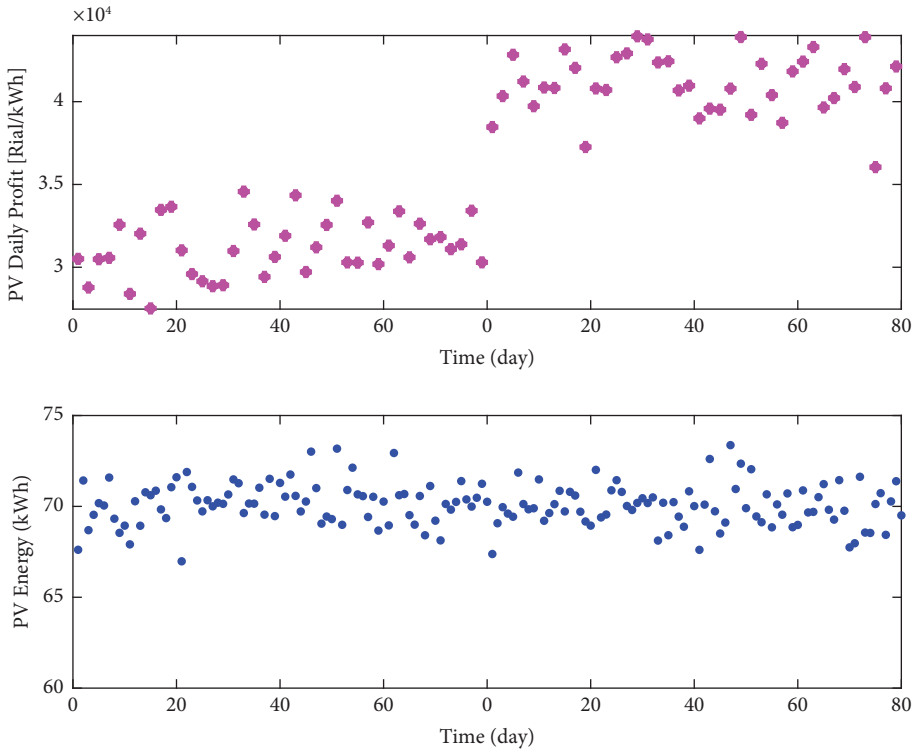

Figure 5: Average daily profit and output power of PV.

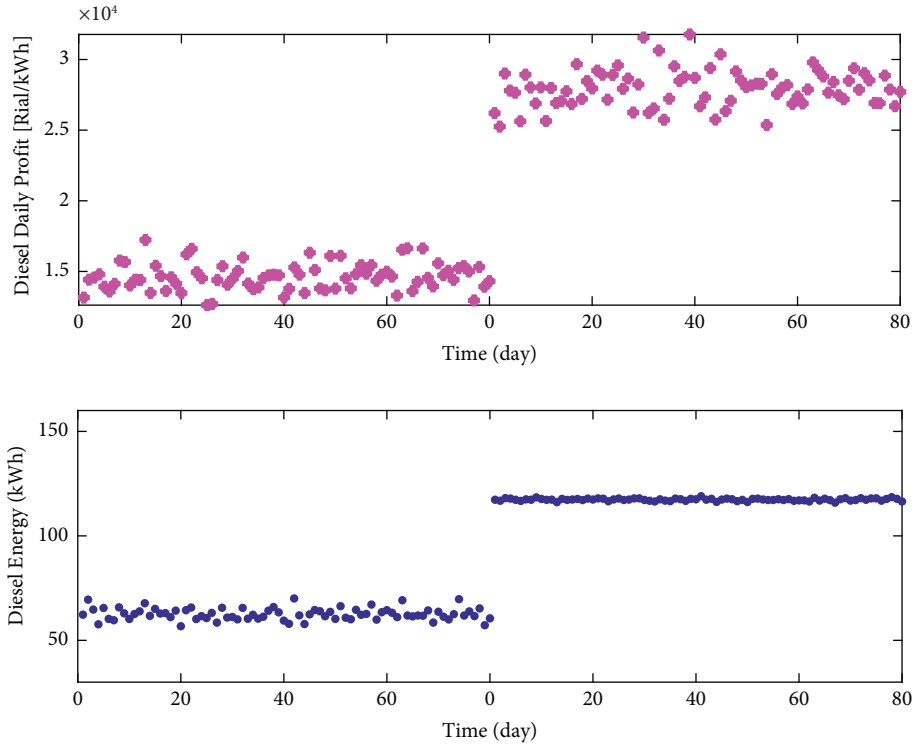

Figure 6: Average daily profit and output power of diesel generator.

The profit of the main grid is also decreased in the second scenario, as illustrated in Figure 12. The profit has become negative. Indeed, the profit from selling energy to the microgrid is lower than the cost of purchasing energy from the microgrid. The power purchased from the main grid is also negative. In other words, the total power received from

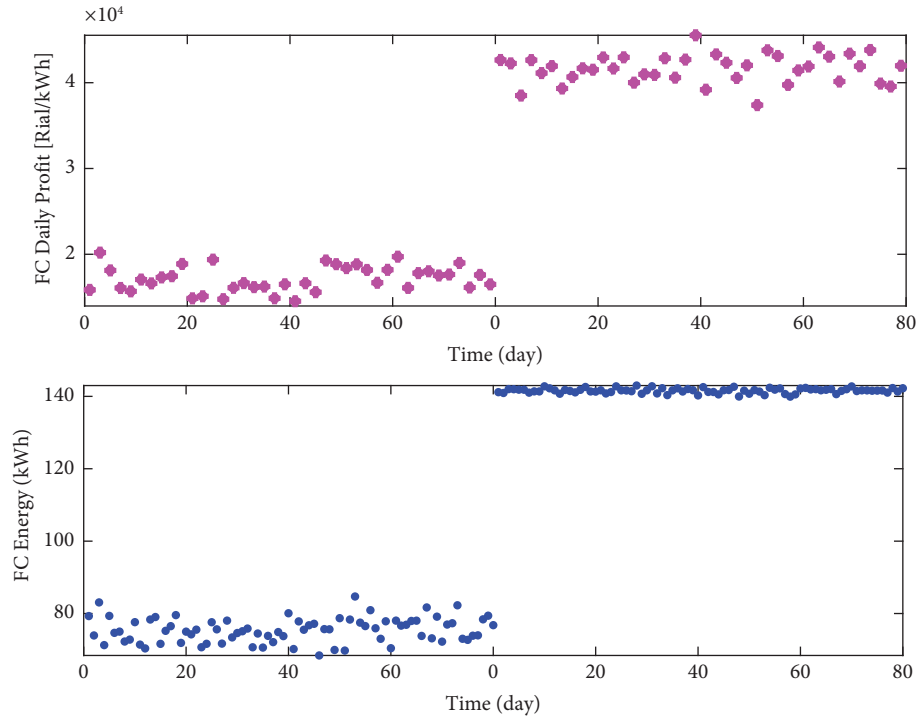

Figure 7: Average daily profit and output power of fuel cell.

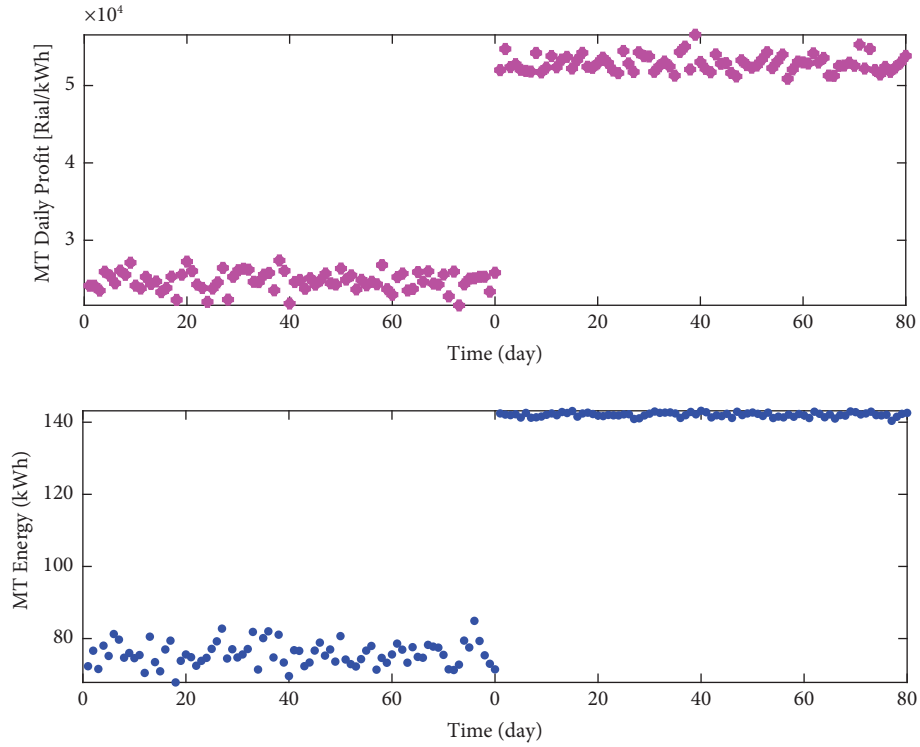

Figure 8: Average daily profit and output power of microturbine.

the main grid is less than the total power given to the main grid. As a result, the dependence of the microgrid on the main grid has been significantly reduced.

Figure 13 shows the hourly amount of profit/cost and consumption/production power of agents. The solar panel can generate energy only during the day from 8 am

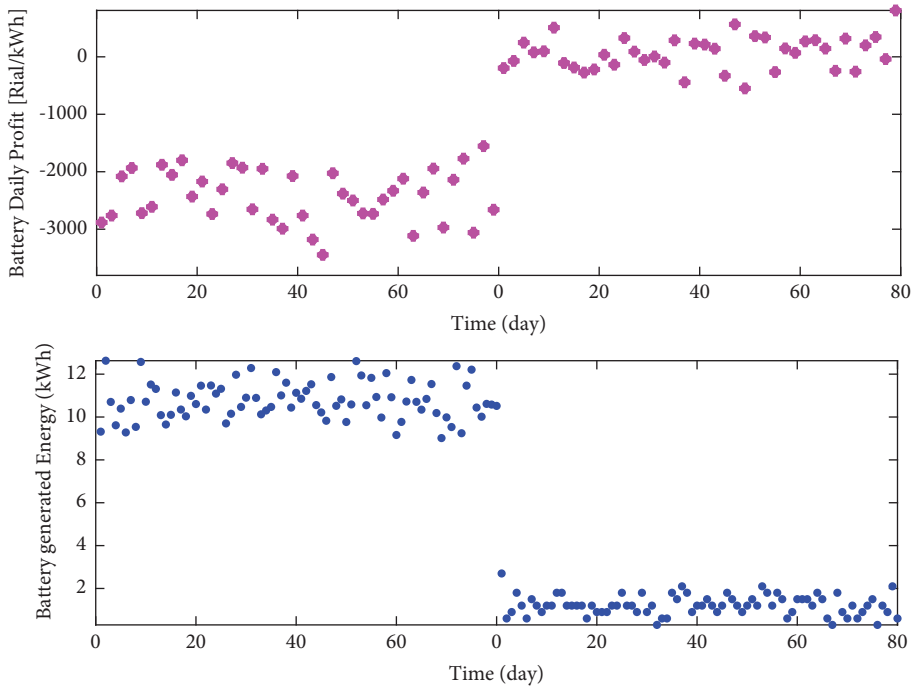

Figure 9: Average daily profit and generated power of the battery.

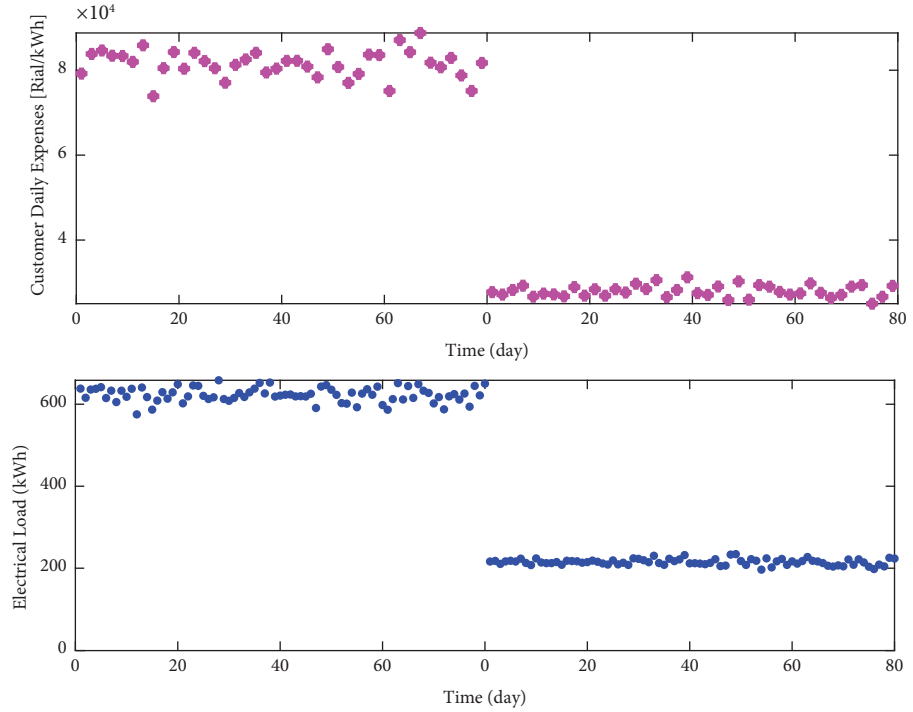

Figure 10: Average daily cost and consumption of the electrical consumer.

to 6 pm in summer. At other times, the output power and profit of the solar panels are zero. The average power output of the wind turbine is almost the same during the day and night hours. Because this graph shows the average output of a wind turbine during 800 days. During the peak hours between 12:00 and 20:00 due to higher demand, the price of energy has increased. Therefore, the profit of wind turbine and other generators including diesel generator, microturbine, and fuel cell has also increased. As expected, the cost and consumed power of consumer agents have also increased during peak consumption hours.

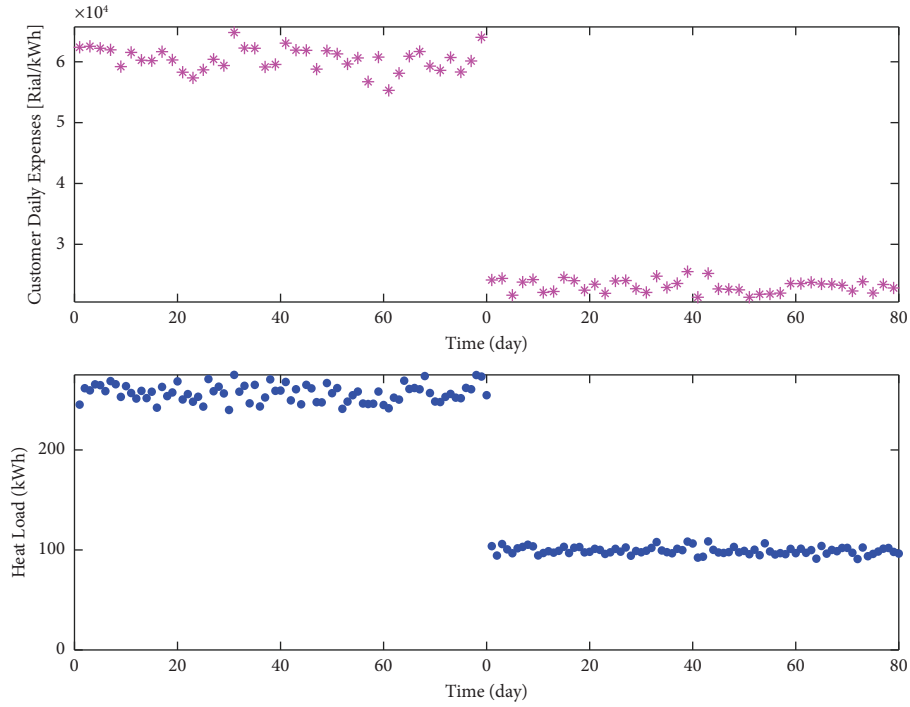

Figure 11: Average daily cost and consumption of the thermal consumer.

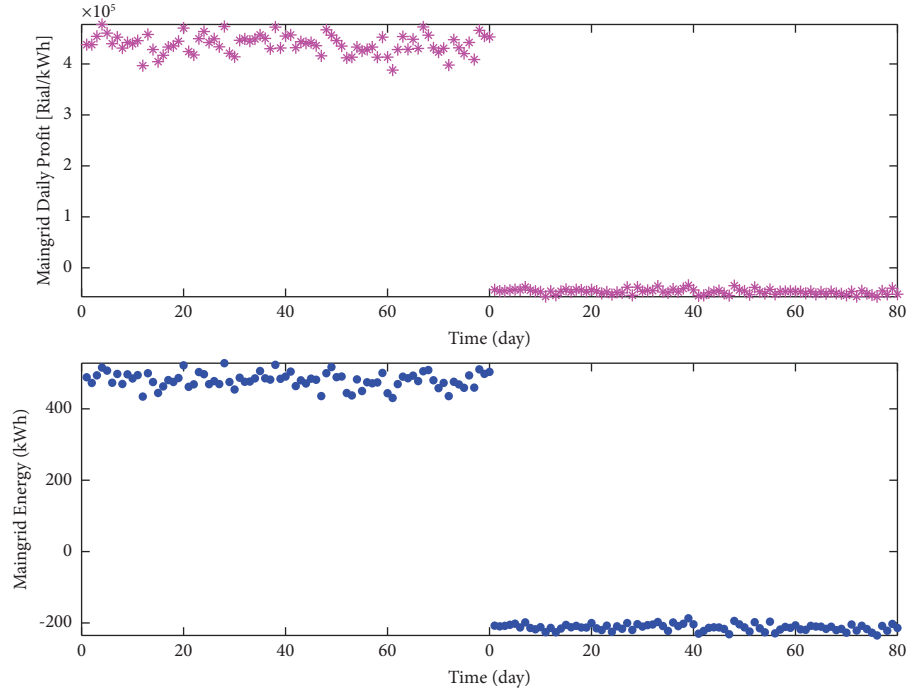

Figure 12: Average daily profit of the main grid and the average daily power received from main grid.

The battery's SOC and energy exchange prices during 24 hours a day are shown in Figure 14. Energy exchange prices include electricity market price, the selling price of energy by the main grid, and the bid price of the battery when it is seller. Electricity market price is actually the price of energy for consumers inside the microgrid and for the battery in purchase mode. According to Figure 14, the battery buys energy when the price of energy is at its lowest. When the price of energy by the main grid is at the highest level, the battery sells energy. Therefore, the battery maximized its profit by offering the highest possible price. It is not economical for the battery to buy/sell the energy immediately when it has just discharged/charged itself. In this case, the battery's lifetime reduces without receiving benefits.

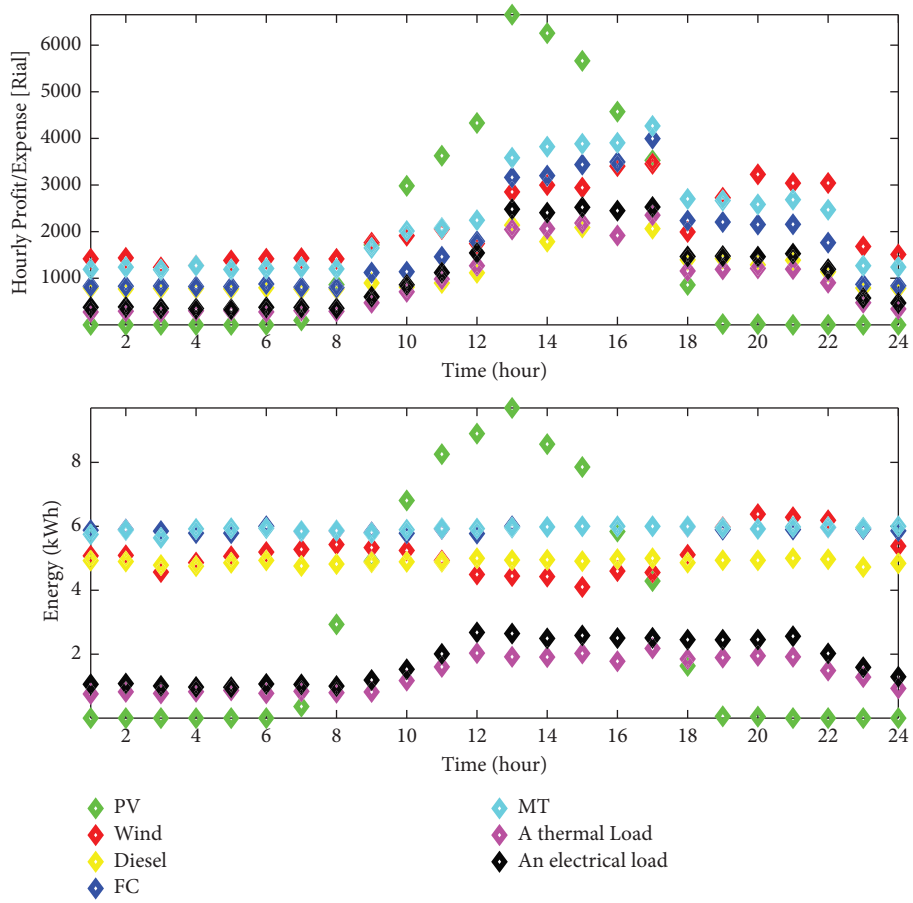

Figure 13: Average hourly profit/cost and production/consumption power of solar panel, wind turbine, diesel, fuel cell, microturbine, and thermal and electrical loads.

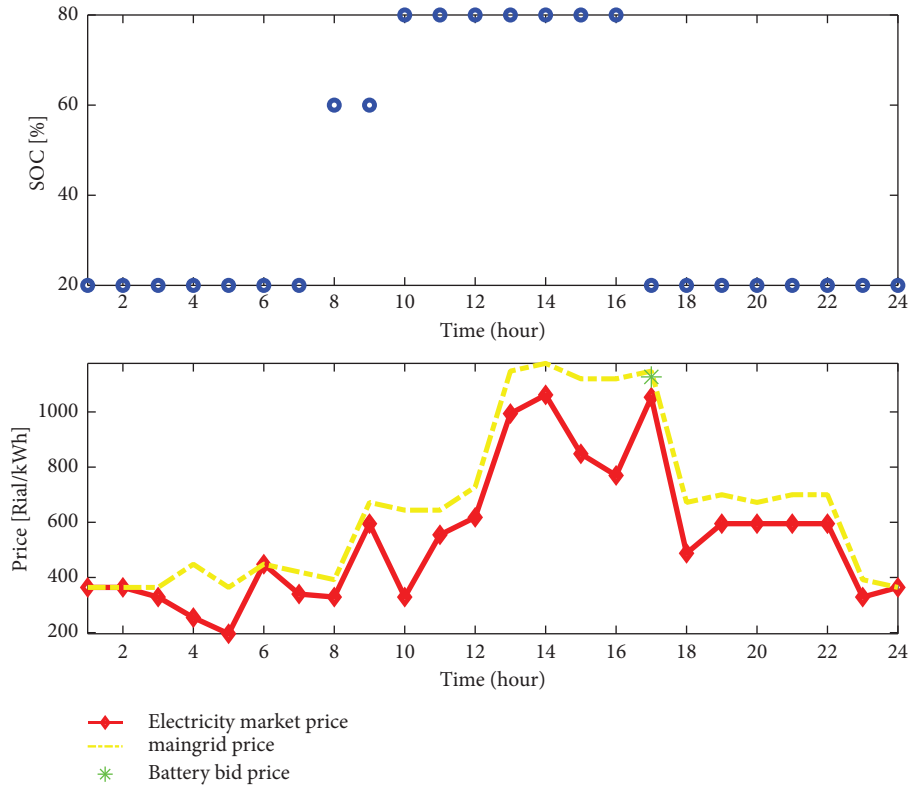

Figure 14: The hourly SOC of the battery and the energy exchange prices.

Table 3: The average results of EMS based on [21] over 800 days.

Scenario	Daily generated/ consumed energy (kWh)		Daily profit/expense (rial)	
	I	II	I	II
Wind	123.4	125.2	40376	49779
PV	70	70	31286	35755
Diesel	63.2	107.8	14834	27385
Fuel cell	75.8	128.6	21145	40953
Microturbine	75.6	130.3	18732	38464
Battery	10.9	2.3	−24161	23.3
Electrical load	620.7	216.4	81437	26642
Thermal load	258.4	99.6	60524	21874
Main grid	491.1	−177.4	441370	−29538

Therefore, after charging/discharge, the battery prefers to be idle for a few hours in order to discharge/charge at the right time. T_k and number of the exploration phase have been selected 180 and 50,000, respectively. Since the number of the states and actions of battery is more than the other agents.

The proposed method is compared with the Q-learning algorithm in [21]. Table 3 shows the simulation results. The profit of generator agents has decreased compared to the proposed method. However, the cost of consumer agents has decreased in the normal Q-learning. For a fair comparison of the two methods, the Fairness Factor (FF) comparison index introduced in [21] was used. In this index, microgrid profit is calculated by considering the profit of both generators and customers. According to Tables 2 and 3, the values of the FF index for the proposed method and normal Q-learning are 1.47 and 1.40, respectively. In both methods, the epsilon value (probability of choosing greedy actions) is zero. The FF index for the method in [21] is significantly smaller than the proposed method. By comparing the FF value in these two methods, it can be concluded that the profit of the microgrid has been significantly improved in the proposed method. Therefore, the decision maker agents of the microgrid have learned to find better policies and the profits of all agents have been maximized, simultaneously. In addition, in the proposed method, the power purchased from main grid is also less than the conventional Q-learning in [21]. Therefore, the dependency of the microgrid on main grid is also reduce in the presented method.

5. Conclusion

This paper presented an intelligent multiagent EMS for a smart microgrid in a stochastic environment. Due to the complexity of the centralized method, the proposed method is an entirely decentralized strategy using reinforcement learning and stochastic games. Unlike many existing methods that use stationary methods for multiagent microgrid energy management, this paper proposed a decentralized strategy compatible with microgrid stochastic structure to manage the hourly electrical and thermal loads of a microgrid. This method converges into an equilibrium solution. Energy resources and consumers were considered as independent and intelligent agents. Agents could learn and can maximize their profits by choosing the right decisions. Reinforcement learning agents discovered the optimal policy by using the feedback from their actions and experiences. Due to the variable property of the output power of RESs and the randomness of the demand, the Markov game has been used to model the random behavior of agents in the microgrid, and the optimal policy of the agents was determined by the decentralized model-free Q-learning. Finally, the results of the proposed method have been compared with the conventional Q-learning algorithm and the satisfactory performance of this method has been shown using actual data of the power grid. In this paper, energy management was considered at the third layer. Implementation of EMS at other layers and adjust the frequency and voltage level of the power grid are suggested as future work. It is also suggested to apply the battery degradation model to increase the efficiency of the battery and its lifetime.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to appreciate the research support of Niroo Research Institute (NRI), Tehran, Iran, and Sun-Air Research Institute of Ferdowsi University, Mashhad, for collecting practical data.

References

[1] V. Vahidinasab, M. Tabarzadi, H. Arasteh et al., "Overview of electric energy distribution networks expansion planning," *IEEE Access*, vol. 8, pp. 34750–34769, 2020.

[2] E. Planas, A. Gil-de-Muro, J. Andreu, I. Kortabarria, and I. Martínez de Alegría, "General aspects, hierarchical controls and droop methods in microgrids: a review," *Renewable and Sustainable Energy Reviews*, vol. 17, pp. 147–159, 2013.

[3] K. Alanne and A. Saari, "Distributed energy generation and sustainable development," *Renewable and Sustainable Energy Reviews*, vol. 10, no. 6, pp. 539–558, 2006.

[4] M. Kia, M. Shafiekhani, H. Arasteh, S. Hashemi, M. Shafie-Khah, and J. Catalão, "Short-term operation of microgrids with thermal and electrical loads under different uncertainties using information gap decision theory," *Energy*, vol. 208, Article ID 118418, 2020.

[5] A. Akbari, V. Vahidinasab, H. Arasteh, and E. Kazemi-Robati, "Rural and residential microgrids: concepts, status quo, model, and application," in *Residential Microgrids and Rural Electrifications*, pp. 131–161, Elsevier, Netherlands, Europe, 2022.

[6] R. S. Sankarkumar and R. Natarajan, "Energy management techniques and topologies suitable for hybrid energy storage system powered electric vehicles: an overview," *International*

Transactions on Electrical Energy Systems, vol. 31, no. 4, Article ID 12819, 2021.

[7] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, "Dynamic energy management system for a smart microgrid," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1643–1656, 2016.

[8] N. Drir, F. Chekired, and D. Rekioua, "An integrated neural network for the dynamic domestic energy management of a solar house," *International Transactions on Electrical Energy Systems*, vol. 31, no. 12, Article ID 13227, 2021.

[9] L. Yang, Q. Sun, D. Ma, and Q. Wei, "Nash Q-learning based equilibrium transfer for integrated energy management game with We-Energy," *Neurocomputing*, vol. 396, pp. 216–223, 2020.

[10] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, no. 11, pp. 1039–1069, 2003.

[11] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, p. 2291, 2019.

[12] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, Article ID 100413, 2021.

[13] J. Jiang, S. Ji, and G. Long, "Decentralized knowledge acquisition for mobile internet applications," *World Wide Web*, vol. 23, no. 5, pp. 2653–2669, 2020.

[14] F.-D. Li, M. Wu, Y. He, and X. Chen, "Optimal control in microgrid using multi-agent reinforcement learning," *ISA Transactions*, vol. 51, no. 6, pp. 743–751, 2012.

[15] T. G. Dietterich, "Hierarchical reinforcement learning with the MAXQ value function decomposition," *Journal of Artificial Intelligence Research*, vol. 13, pp. 227–303, 2000.

[16] Z. Hu, M. Zhu, P. Chen, and P. Liu, "On convergence rates of game theoretic reinforcement learning algorithms," *Automatica*, vol. 104, pp. 90–101, 2019.

[17] C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li, "Multi-UAV trajectory planning for energy-efficient content coverage: a decentralized learning-based approach," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3193–3207, 2021.

[18] V. Boglou, C. S. Karavas, A. Karlis, and K. Arvanitis, "An intelligent decentralized energy management strategy for the optimal electric vehicles' charging in low-voltage islanded microgrids," *International Journal of Energy Research*, vol. 46, no. 3, pp. 2988–3016, 2022.

[19] C.-S. Karavas, G. Kyriakarakos, K. G. Arvanitis, and G. Papadakis, "A multi-agent decentralized energy management system based on distributed intelligence for the design and control of autonomous polygeneration microgrids," *Energy Conversion and Management*, vol. 103, pp. 166–179, 2015.

[20] C.-S. Karavas, K. Arvanitis, and G. Papadakis, "A game theory approach to multi-agent decentralized energy management of autonomous polygeneration microgrids," *Energies*, vol. 10, no. 11, p. 1756, 2017.

[21] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5749–5758, 2018.

[22] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020.

[23] D. S. Leslie and E. J. Collins, "Individual Q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.

[24] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings*, pp. 157–163, Elsevier, Netherlands, Europe, 1994.

[25] S. Hashemi, H. Arasteh, M. Shafiekhani, M. Kia, and J. Guerrero, "Multi-objective operation of microgrids based on electrical and thermal flexibility metrics using the NNC and IGDT methods," *International Journal of Electrical Power & Energy Systems*, vol. 144, Article ID 108617, 2023.

[26] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, 2018.

[27] B. Belousov, H. Abdulsamad, P. Klink, S. Parisi, and J. Peters, *Reinforcement Learning Algorithms: Analysis and Applications*, Springer, Heidelberg, Germany, 2021.

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT press, Cambridge, UK, 2018.

[29] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3/4, pp. 279–292, 1992.

[30] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.

[31] A. M. Fink, "Equilibrium in a stochastic $ n $-person game," *Hiroshima Mathematical Journal*, vol. 28, no. 1, pp. 89–93, 1964.

[32] V. Vahidinasab, "Optimal distributed energy resources planning in a competitive electricity market: multiobjective optimization and probabilistic design," *Renewable Energy*, vol. 66, pp. 354–363, 2014.

[33] R. Darshi, M. A. Bahreini, and S. A. Ebrahim, "Prediction of short-term electricity consumption by artificial neural networks levenberg-marquardt algorithm in hormozgan province, Iran," in *Proceedings of the 2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 1–4, IEEE, Shahrood, Iran, April 2019.